

PipeGen: Data Pipe Generator for Hybrid Analytics

Brandon Haynes

Alvin Cheung

Magdalena Balazinska

Department of Computer Science & Engineering
University of Washington
{bhaynes, akcheung, magda}@cs.washington.edu

ABSTRACT

We develop a tool called PipeGen for efficient data transfer between database management systems (DBMSs). PipeGen targets data analytics workloads on shared-nothing engines. It supports scenarios where users seek to perform different parts of an analysis in different DBMSs or want to combine and analyze data stored in different systems. The systems may be co-located in the same cluster or may be in different clusters. To achieve high performance, PipeGen leverages the ability of all DBMSs to export, possibly in parallel, data into a common data format, such as CSV or JSON. It automatically extends these import and export functions with efficient binary data transfer capabilities that avoid materializing the transmitted data on the file system. We implement a prototype of PipeGen and evaluate it by automatically generating data pipes between five different DBMSs. Our experiments show that PipeGen delivers speedups up to $3.8\times$ compared with manually exporting and importing data across systems using CSV.

1. INTRODUCTION

Modern data analytics often requires retrieving data stored in multiple database management systems (DBMSs). As an example, consider a scientist who collects network data in a graph database such as Giraph [7], while storing other metadata in a relational database system such as Derby [4]. To analyze her data, the scientist must retrieve data from the relational store, and subsequently combine it with the data stored in the graph database to produce the final result. As part of the process, she might further need to temporarily move the retrieved data to other DBMSs to leverage their specialized capabilities such as machine learning algorithms [11], array processing operators [39], and so on. This type of federated data analysis is referred to as *hybrid analytics* and remains poorly supported today even as the landscape of big data management and analytics systems is rapidly expanding.

In general, a hybrid analytics task involves moving data among n different DBMSs, with each DBMS potentially having its own internal data format. A critical requirement of hybrid analytics is the ability to move data efficiently among the different DBMSs involved [37]. There have been three main approaches to solving this

problem. The simplest approach is to use an intermediate extract transform and load (ETL) process that issues queries to the source DBMS using protocols such as JDBC [2] or through a specialized wrapper, extracts the data, then inserts it into the target DBMS. This approach, however, fails to leverage the fact that the source and target systems may both be shared-nothing and are likely executing in the same cluster. It is thus ill-suited for moving large datasets.

As a faster approach, many DBMSs provide means for users to export the stored data, often in parallel, to an external format, and import the data from that same format. The user can leverage such a capability to transfer data between the two DBMSs by first exporting data from the source system, and subsequently importing it into the target DBMS. This works particularly well in settings where a common data format exists between the source and target DBMSs, for instance text-based data formats such as comma-separated values (CSV) or binary data formats such as Protocol Buffers [32], Parquet [31], and Arrow [6]. Unfortunately, moving data using a text-oriented format such as CSV or JSON is extremely costly: it requires serializing the data to be transferred from the internal format of the source DBMS to the common text format, storing the serialized data to physical storage, and finally importing from storage into the internal format of the target DBMS.¹ Shared binary formats alleviate some overheads but still require transforming and materializing the transferred data, which incurs unnecessary and non-trivial cost in both time and storage space. Additionally, while support for text-oriented formats such as CSV are common, shared binary formats remain rare and evolve rapidly. As a concrete example, CSV is the only common data format supported by the five DBMSs (Myria [20], Spark [44], Giraph [7], Hadoop [5], and Derby [4]) that we use in the evaluation.

A third approach to moving data that avoids using physical storage as intermediary is to construct dedicated data transfer programs, i.e., *data pipes*, between specific source and target DBMSs. This ranges from implementations of common data transfer protocols such as JDBC [2], ODBC [30], and Thrift [41], to software packages for transferring data between two specific systems, such as spark-sframe for moving data between Spark and GraphLab [11]. Unfortunately, generalizing this approach requires implementing $O(n^2)$ data pipes in order to transfer data between n different DBMSs. In addition to being impractical, this approach requires knowledge about the internals of the source and target DBMSs, making it inaccessible to non-technical users. Even for technical professionals, implementing dedicated data pipes is often an error-prone process; the pipe implementations are often brittle

¹In this paper, we use DBMS to refer to both relational and non-relational stores. Also, we assume the transferred data will be discarded and not persisted in the target DBMS after the user has examined the query results.

as systems evolve over time and this renders previous pipe implementations obsolete.

In this paper, we describe a new tool called PipeGen that retains the benefits of dedicated data pipes without the shortcomings of manual implementation or serializing via physical storage. Using a combination of static and dynamic code analysis, PipeGen automatically constructs and embeds a data pipe into a DBMS by exploiting the fact that many DBMSs can export and import data from commonly-used formats (e.g., CSV), and that most systems come with unit tests that exercise that functionality. PipeGen takes a number of inputs, including a set of such unit tests and the source code of the DBMS. PipeGen analyzes the source code of the DBMS and executes each of the export unit tests to create a data pipe that redirects data being exported to the disk to a network socket that is provided by PipeGen at runtime. Similarly, the import unit tests are used to create a data pipe that reads data from a network socket rather than a disk file. The same unit tests are used to validate the correctness of the generated data pipes. Users then leverage the generated pipes by issuing queries that export and import from disk files just like before, except that PipeGen connects the generated pipes together via sockets so that data can be transmitted directly from the source to the target DBMS, bypassing the disk, and leveraging parallelism when possible.

PipeGen comes with a number of optimizations to improve the performance of the generated data pipes. In particular, PipeGen extends the text-oriented data format implementations with extra code that *intercepts the original data* before its text encoding, identifies and eliminates delimiters, and identifies and eliminates redundant metadata. By capturing the original data, PipeGen enables the transmission of that data using an efficient binary format (Apache Arrow [6] in our prototype implementation). PipeGen thus uses existing CSV or JSON data export and import capabilities as a specification to automatically implement an efficient, binary data pipe. With this approach, PipeGen not only enables efficient binary data transfers but also simplifies the evolution of the binary data transfer formats: whenever a new format becomes available (e.g., the next generation of Apache Arrow), it suffices to update the PipeGen tool and that new format automatically becomes available to transfer data between different big data systems, without having to implement support for the new format in each of these systems separately.

Our experiments show that PipeGen can generate data pipes that speed up data transfer between DBMSs by factors up to $3.8\times$, both when the source and target systems are colocated on a single node, and when they are placed on separate machines.

PipeGen currently does not address the problem of schema matching and focuses on the mechanics of data movement instead. We expect the generated data pipes to be utilized directly by the end user or a query optimizer, which would insert additional operators to perform schema matching either before or after data has been transferred.

In summary, this paper makes the following contributions:

- We describe an approach based on program analysis to automatically create data pipes and enable direct data transfer between DBMSs (Section 3 and Section 4).
- We present a number of techniques that improve the performance of the generated data pipes by removing the unnecessary computation that are inherent in text-oriented formats (Section 5). Importantly, these optimizations enable PipeGen to automatically replace an inefficient text-based data pipe with an efficient, binary one. Through this approach, PipeGen makes fast data transfer formats available to

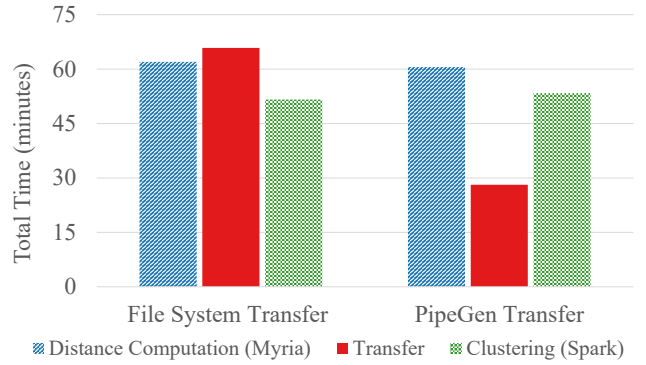


Figure 1: Total time for an astronomy workflow with and without PipeGen-generated data pipes. Pairwise distance between each tuple pair was computed in Myria and transferred to Spark where PIC clustering [27] was performed.

data analytics engines without requiring developers to manually change the source code of these engines.

- We build a prototype of PipeGen, and evaluate it by generating data pipes between five different Java-based DBMSs. The optimized data pipes are up to $3.8\times$ faster than a naïve approach that transfers data via the file system using CSV.

2. MOTIVATING EXAMPLE

Consider an astronomer who studies the evolution of galaxies using N-body simulations of the universe [23]. A typical workflow for this analysis is for the astronomer to first collect the output of a cosmological simulation. The output takes the form of a series of snapshots of the universe. Each snapshot consists of particles distributed in 3D space. The astronomer then needs to cluster the particles for each snapshot into galaxies in order to analyze the evolution of these galaxies over time. In this example, the particle clustering is a critical piece of the analysis and astronomers often experiment with different clustering algorithms [25].

In our scenario, an astronomer uses the Myria DBMS [20, 29] to analyze the output of an N-body simulation, including performing an initial particle clustering, when she learns of a novel clustering method that may be of interest. This new method may or may not outperform the current approach, and the astronomer is interested in evaluating the two. Critically, however, Myria *does not support this technique*, but it is available in Spark [44].

A typical approach involves the following: perform some or all of the data preparation tasks in Myria, export the intermediate result to the file system, and import the files into Spark. Since the only common file format supported by both systems is CSV (both also support JSON but produce incompatible documents), the intermediate result is materialized as a set of CSV files. A second iteration is necessary to bring results back to Myria for comparison.

Unfortunately, exporting and importing large data via the file system is an expensive proposition, and for some datasets there may be insufficient space available for a second materialization. The alternate approach of modifying the DBMS source code to support efficient data transfer requires deep programming and database systems expertise. Unfortunately, this data movement challenge makes it very difficult or impossible for domain scientists to leverage the highly-specialized functionalities across different DBMSs.

The PipeGen tool avoids these disadvantages. To illustrate, we execute the workflow described above on the present-day snapshot of a 5TB astronomy simulation stored in the Myria. The present day snapshot is 100 GB in size and an initial data clustering has

already been performed on the data. Our goal is to count differences in cluster assignments between the existing data clustering and those for power iteration clustering (PIC) [27] available on Spark. We compute pairwise distances in Myria between all particles having distance less than threshold $\epsilon = 0.00024$, set by our astronomy collaborators. This yields approximately one billion pairs. We then transfer this data to Spark where we perform PIC clustering. Finally, we transfer the resulting assignments back to Myria for comparison with the existing clusters. To contrast different data transfer mechanisms, we perform the transfer using both the file system as intermediary and data pipes generated by PipeGen.

Figure 1 illustrates the performance of each of these steps. When using the file system, transfer of the large set of intermediate pairwise differences is more expensive than either of the other two steps alone. Using the data pipe generated by PipeGen, however, reduces the time spent in data transfer from 66 to 28 minutes. This result suggests the importance of having efficient data transfers in support of hybrid data analytics, and we explain how PipeGen generates data pipes automatically in the following sections.

3. THE PIPEGEN APPROACH

PipeGen enables users to move data efficiently between DBMSs. In this section, we give an overview of PipeGen’s usage and approach.

3.1 Using PipeGen

To use data pipes, a user first invokes PipeGen to generate the source code for data pipes. After compilation, users can write queries to leverage the generated pipes. Below we give an overview of the two steps.

Constructing a data pipe. To generate a data pipe, the user invokes PipeGen with a number of inputs. These include (1) a script used to build the DBMS along with its source code, (2) a pair of scripts that execute unit tests associated with import and export functionality for a specific data format (e.g., CSV, JSON, or binary format), (3) the name of the specific data format produced during import and export (for use during library substitution; see Section 5.2), (4) additional configuration-related metadata.

Given these inputs, PipeGen analyzes the source code of the DBMS, and generates a data pipe that exports data via to the network socket to be instantiated by PipeGen during runtime.

Using the generated data pipe. To use the data pipe generated by PipeGen, the user executes two queries: one that exports data from the source DBMS, and one that imports data into the target DBMS, as shown in Figure 2. These queries may occur in any order; PipeGen will automatically block until both DBMSs are ready (see Section 4.2). Note that the queries issued to each DBMS are written as if data are moved from the source to target DBMS via physical storage: users only need to specify the data to be exported from the source and the name of the target DBMS using the special “db://X” syntax rather than a filename.

After receiving the queries, PipeGen connects the generated data pipes in the source and target DBMSs by passing them a network socket to transmit data. In the case where the source and target DBMSs support multiple worker threads, they are matched with each other by the worker directory maintained by PipeGen, as shown in Figure 2.

Usage in the context of hybrid and federated systems. Because of the flexibility that PipeGen’s approach affords, we also expect that our automatically-generated data pipes are well-suited for use in other contexts such as by being integrated into an existing hybrid (e.g., [14, 26]) or federated DBMS (e.g., Garlic [24]).

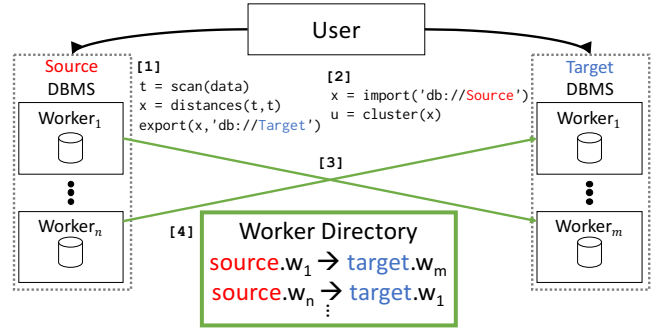


Figure 2: Using the data pipe generated by PipeGen for the hybrid analysis from Section 2: 1. User submits a query to the source DBMS (e.g., Myria) to compute distance and export to the target DBMS using data pipe. 2. User issues an import query on the target DBMS (e.g., Spark) to cluster the result. Data is transferred using the generated data pipe in 3., and in 4. a worker directory (see Section 4.2) coordinates the connection process. PipeGen-related components are highlighted in green.

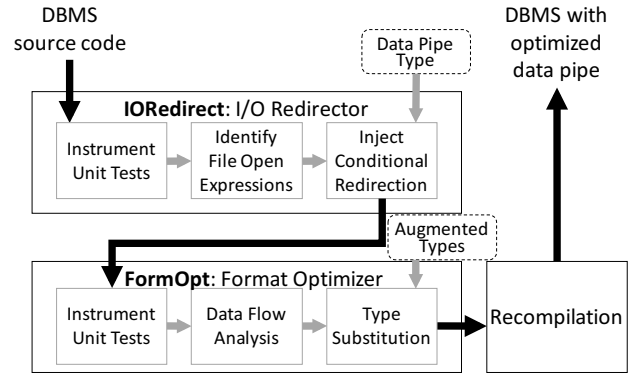


Figure 3: Compile-time components of PipeGen. The file IO redirector (IORedirect) generates a data pipe to transfer data via a network socket, while the format optimizer (FormOpt) improves the efficiency of data transfer for text-oriented formats.

3.2 Data Pipe Generation Overview

To generate a data pipe, PipeGen takes as input the source code of a DBMS and tests that exercise the import and export functionality for a given format. It begins by executing and analyzing the provided tests. Using the results, it modifies the source code of the DBMS to generate a data pipe that can transfer data directly between the DBMSs using a common data format. When the common format is text-oriented, PipeGen further optimizes the format of the transferred data to improve performance, as we will discuss in Section 5.

PipeGen supports single-node and parallel DBMSs. For the latter, as illustrated in Figure 2, the source code produced transfers data in parallel directly between individual workers. Our implementation and evaluation currently target shared-nothing systems, but the approach can be applied to other parallel architectures.

To generate a data pipe, PipeGen performs the two-phase compilation process shown in Figure 3. First, PipeGen locates in the DBMS’s source code the relevant code fragments that import from and export data to disk files. For each of these fragments, PipeGen identifies the file system operations (e.g., file open and close) and substitutes them with equivalents on a network socket. It accomplishes this by instrumenting the provided unit tests and identify-

ing the expressions in the code that open the import or export file on disk. It then modifies these expressions to allow direct sending and receiving of data via a network socket instead of going through the file system. This approach enables datasets of arbitrary size to be transmitted (in contrast to an approach based on memory-mapped files, for example), and for file systems that do not support named pipes (e.g., HDFS). To further improve the performance of the generated data pipe, the format optimizer eliminates unnecessary operations for text-oriented formats. We respectively explain both phases in detail in the following sections.

Assumptions. PipeGen assumes that the implementation of a given DBMS’s data import and export is well-behaved. For instance, PipeGen requires that the implementation not perform large seeks during file IO and that an import or export file is opened exactly once, since otherwise multiple sockets will be created during pipe generation. For text-based formats, PipeGen assumes that character strings are used to serialize data (rather than byte arrays, for instance), and that values are constructed through string concatenation rather than random access. PipeGen also assumes that, for a given transfer, each DBMS support the same data format. Finally, PipeGen assumes that the unit tests fully exercise the code associated with the import and export of data. For export, this includes all code paths between reading data in the internal representation and serializing them to disk using system calls. For import this includes reading serialized data from disk back into the internal data representation.

4. FILE IO REDIRECTION

In this section, we discuss the redirection component of PipeGen. This component, called *IORedirect*, creates a data pipe from the DBMS’s existing serialization code and modifies that code such that instead of using the file system, data are exported to (and imported from) a network socket provided by PipeGen at runtime. When the source and target DBMSs are colocated on the same machine, the socket is simply a local loopback one. Otherwise the socket connects to the remote machine hosting the target DBMS, with the address provided by a directory (Section 4.2).

4.1 Basic Operations

In order to modify a DBMS’s source code in this way, *IORedirect* first identifies the relevant file system call sites that open the disk file used for import or export. It does not suffice to search the source code for “open file” operations, however, because a DBMS may open multiple, unrelated files such as debugging logs. *IORedirect* should not inadvertently replace such files with sockets.

IORedirect disambiguates this by instrumenting all file open calls in the source code, executing the import and export unit tests, and capturing the filename that is passed in to each call. All calls with filenames other than the target of the import / export are eliminated. *IORedirect* then modifies each remaining call site by adding code that redirects I/O operations to a network socket. The new code is executed only when a user specifies a *reserved filename* for import or export. The conditional nature of the redirection is important because we want to preserve DBMS’s ability to import from and export to disk files. This augmented architecture is illustrated in Figure 4.

When adding the conditional logic to each relevant call site, *IORedirect* introduces a specialized data pipe class into the DBMS’s source code that reads or writes to the remote DBMS via a passed-in network socket rather than the file system, but is otherwise a language-specific subtype of its file system-oriented counterpart. This allows PipeGen to substitute all instances of the file IO classes with a data pipe instance. To illustrate, a Berkeley socket

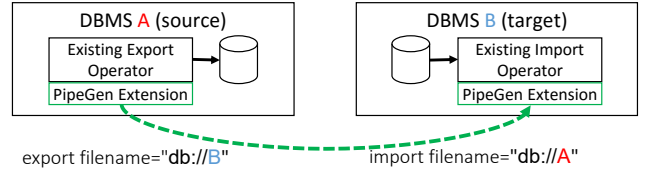


Figure 4: PipeGen creates a data pipe to transmit data via a network socket. User activates the data pipe by specifying a reserved filename as the export and import destination.

```
1 void export(String filename, ...) {
2     ...
3     stream = new FileOutputStream(filename);
4     ...
5     stream.close();
6 }
```



```
1 void export(String filename, ...) {
2     ...
3     stream = filename.matches(format)
4         ? new DataPipeOutputStream(filename);
5       : new FileOutputStream(filename);
6     ...
7     stream.close();
8 }
```

Figure 5: An example modification to create a Java data pipe. Here the unit tests indicate line 3 (top) to be modified. The modified code below, matches against the reserved filename format (bottom, line 3), and uses the generated data pipe if a match is found (bottom, line 4). Due to subtyping, subsequent uses of the resulting stream are unaffected (e.g., line 7).

descriptor can be substituted for a file descriptor in a C or C++-like language, while Java or .NET file streams can be interchanged with network socket streams.

IORedirect modifies the code to use the generated data pipe when the reserved filename is specified. Figure 5 illustrates one such modification made by *IORedirect* for the Myria DBMS, which is written in Java. Here an instantiation of `FileOutputStream` is replaced with a ternary operator that examines the filename and substitutes a data pipe class (itself a subtype of `FileOutputStream`) upon detecting a match. Due to subtyping, other IO operations (e.g., `close`) operate as before.

PipeGen verifies the correctness of the generated data pipe as follows. First, it launches a *verification proxy* that imports and exports data as if it were a remote DBMS. This proxy redirects all data received over a data pipe to the file system, and transmits data from the file system through a data pipe. Next, the unit tests for the modified DBMS are executed using a reserved filename format that activates the data pipe for all files. As data are imported and exported over the data pipe, the proxy reads and writes to disk as if the data pipe did not exist. We then rely on existing unit test logic to verify that the contents that are read from or written to disk are correct.

Finally, *IORedirect* exposes a dynamic debugging mode that probabilistically verifies the correctness of the generated data pipe at runtime, and may be used when the source and destination DBMSs are colocated. Under this mode, the first n elements of the transmitted data are both written to the disk using existing serialization logic, and transmitted over the data pipe. The receiving DBMS then reads the data from the file system and compares it to the values that have been transmitted over the pipe. Any deviations trigger a query failure.

```

1 class DataPipeOutputStream extends
    FileOutputStream {
2     DataPipeOutputStream(String fname) {
3         e = Directory.query(fname);
4         socket = new Socket(e.hostname, e.port);
5         ...
6     } }

```

Figure 6: Simplified Java implementation of directory querying logic. The query (line 3) blocks until an import worker has registered its details with the directory. The resulting directory entry is used to open a socket to the remote DBMS (line 4).

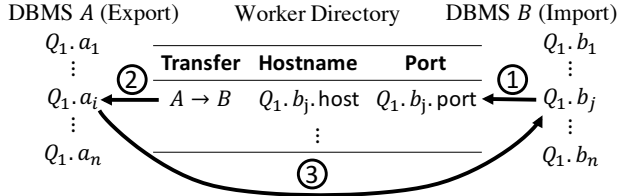


Figure 7: Using the worker directory to initiate data transfer. In 1., import worker b_j from query Q_1 registers with the directory and awaits a connection. In 2., export worker a_i queries the directory and blocks until an entry is available, after which it uses the associated details to connect in 3.

4.2 Parallel Data Pipes

Many multi-node DBMSs support importing or exporting data in parallel using multiple worker threads. PipeGen uses a *worker directory* to match up the worker threads in the source and target DBMSs. The directory is instantiated by PipeGen when the DBMS starts and is accessible by all DBMS worker processes.

The worker directory is used as follows. When a user exports data from DBMS A to DBMS B, as DBMS B prepares to import, each worker b_1, \dots, b_n registers with the directory to receive data from DBMS A. As data is exported from DBMS A, workers a_1, \dots, a_n query the directory to obtain the address and port of a receiving worker b_i . Each exporting worker a_i blocks until an entry is available in the directory, after which it connects to its corresponding worker using a network socket. To allow multiple concurrent transfers between the same pair of DBMSs, each import and export query is also assigned a unique identifier to disambiguate queries executing simultaneously. The data pipe class performs the directory registration or query process. For example, a simplified Java implementation for the query process is shown in Figure 6, and Figure 7 describes the overall workflow.

The worker directory ordinarily assumes that the number of workers between the source and target DBMSs are identical. However, a user may embed metadata to explicitly indicate the number of exporting and importing processes. For example, an export from DBMS A using two workers to DBMS B with three workers would use the respective reserved filenames `db://A?workers=2` and `db://A?workers=3`. When the number of importing workers exceeds the exporters, the worker directory opens a “stub” socket for the orphaned importing worker that immediately signals an end-of-file condition; under this approach the extra importing workers simply sit idle until the data transfer has completed. The case where the number of exporting workers exceeds the importing workers is left as a future extension.

5. OPTIMIZATIONS

When a pair of DBMSs share an optimized binary format, the data pipe that IORedirect generates can immediately be used to

transfer data between the two systems. For example, PipeGen can generate a data pipe that transfers data between Spark and Giraph using Hadoop sequence files, as that format is supported by both systems. Unfortunately, it is rare for two systems to support the same efficient binary format. For example, the Apache Parquet format is (natively) supported by only two of the systems we evaluate (Spark [44] and Hadoop [5]). Giraph does not support Parquet without third-party extension while Apache Derby [4] does not support import and export of any binary format.

By contrast, support for text-oriented formats is far more pervasive: *all* of the DBMSs we evaluate (Myria [20], Spark [44], Giraph [7], Hadoop [5], and Derby [4]) support bulk import and export using a CSV format, and all but Derby support JSON. Accordingly, we expect users to frequently revert to text-based data formats to transfer data between such systems.

Unfortunately, text-based formats are inefficient and incur substantial performance overhead. The main sources of this include:

- **String encoding of numeric types.** It is often the case that the size of a converted string greatly exceeds that of its original value, which increases transfer time. For example, the 4-byte floating point value $-2.2250738585072020 \cdot 10^{-308}$ requires 24 bytes to encode as a string.
- **String conversion and parsing overhead.** Text-oriented export and import logic spends a substantial amount of time converting primitive values into/from their string representation.
- **Extraneous delimiters.** For attributes with fixed-width representations, many data formats include extraneous separators such as value delimiters and newlines.
- **Row-orientation.** DBMSs that output CSV and JSON often do so in a row-oriented manner; this is the case for all of the DBMSs we evaluate in Section 7. This makes it difficult to apply layout and compression techniques that benefit from a column-oriented layout. For instance, our experiments show that converting intermediate data to column-major form offers a modest performance benefit (see Figure 12).

To improve the performance of text-based formats, PipeGen comes with a format optimizer called *FormOpt* (see Figure 3), to address the inefficiencies listed above. FormOpt optimizes a given text format in one of two modes. First, it analyzes the DBMS source code to determine if the export or import logic use an external library to serialize data to JSON or CSV (PipeGen currently supports the Jackson JSON library [22] and we plan to support Java JSONArrays and the Apache Commons library [3]). If so, then FormOpt replaces use of the library with a PipeGen-aware variant. The implementation of this variant is then directly responsible for performing the optimizations described above.

On the other hand, a DBMS might directly implement serialization functionality instead of using an external library. For example, the Myria DBMS directly implements its JSON export functionality. For such systems, FormOpt supports a second mode that leverages *string decoration* to target the actual string production and consumption logic that takes place during data export and import. Since components of the string decoration mode are used when applying the library extension strategy, we introduce it first and then turn to the library extension process.

5.1 String Decorations

Without using external libraries, we assume the DBMS uses the following string operations to serialize data into CSV or JSON text:

convert objects and primitive values into strings, concatenate the strings, intersperse delimiters and metadata such as attribute names, and write the result out. To optimize these steps, FormOpt modifies the DBMS source code such that, whenever the modified DBMS attempts to write the string representation of a fixed-width primitive to the stream, it instead writes to a compact binary representation provided by PipeGen. Ordinary strings and other non-primitive values are transmitted in their unmodified form. As we will see, doing so eliminates the transmission of unnecessary values such as delimiters and attribute names.

To accomplish this source code modification, PipeGen’s data pipe type (introduced in Section 4.1) must receive the primitive values of the data to transmit before they are converted into strings. The core difficulty in ensuring this property is that there may be many primitive values embedded in any given string that is passed to the data pipe type. For example, a particular DBMS might concatenate together all attributes before writing them to the output stream `s.write(attr1.toString() + “,” + attr2.toString() + ...)`. By the time the data pipe type receives the concatenated value, it will be too late as both of the attributes would have already been converted into strings.

FormOpt addresses this by introducing a new augmented string called `AString`, which is a subtype of `java.lang.String`. `AString` is backed by an array of Objects rather than characters. By substituting `java.lang.String` instances for `AStrings` in the right locations, FormOpt avoids the problem described above by storing references to the objects to be serialized rather than their string representation.

For example, given ordinary string concatenation:

```
s = new Integer(1).toString() + "," + "a";
```

FormOpt changes the statement into one that uses `AStrings`:

```
s = new AString(1) + new AString(",") +
    new AString("a");
```

Each of the three instances maintains its associated value as an internal field (1, “,” and “a” respectively) and the concatenated result—itsself an `AString` instance—internally maintains the state [1, “,”, “a”]. Note that the final `AString` instance need not include the concatenated string “1,a” in its internal state since it may easily reproduce (and memoize) it on demand. More complex types are immediately converted into strings during this aggregation process to ensure that subsequent changes to their state do not affect the internal state of the `AString` instance. However, as we shall see below, converting a complex object into a string (e.g., through a `toString` invocation) may produce an `AString` instance, which allows for nesting under supported formats.

When an IO method is invoked on the data pipe type, its implementation inspects any string parameter to see if it is an `AString`. Additionally, methods exposed by the data pipe type that produce a string return an `AString`. During export, this allows the data pipe type to directly utilize the unconverted values present in the internal state of an `AString`; similarly, during import the `AString` implementation efficiently executes common operations such as splitting on a delimiter and conversion to numeric values without materializing as character string.

This resolves the problem we described above, but does not address the issue of *where* to substitute an `AString` for a regular string instance. Intuitively, we want to substitute only values that are (directly or indirectly) related to data pipe operations, rather than replacing all string instances in the source code. To find this subset, PipeGen executes each of the provided unit tests and marks all call sites where data is written to/read from the data pipe. PipeGen then performs data-flow analysis to identify the sources of those values (for export) and conversions to primitive values (for import). This

Algorithm 1 String decoration

```
function TRANSFORM( $T$ : tests)
1: for each  $t \in T$  do
2:   Find relevant IO call sites  $C$ 
3:   for each  $c \in C$  do
4:     Construct data-flow graph  $G$ 
5:     for each expression  $e \in G$  do
6:       if  $e$  is literal  $v$  or
7:          $e$  is instantiation String( $v$ ) or
8:          $e$  is v.toString() then
9:         Replace  $v$  with AString( $v$ )
10:      else if  $e$  is Integer.parseInt( $v$ ) then
11:        Replace  $v$  with AString.parseInt( $v$ )
12:      else if  $e$  is Float.parseFloat( $v$ ) then
13:        Replace  $v$  with AString.parseFloat( $v$ )
14:      Similarly for other string operations
```

produces a data-flow graph (DFG) that identifies candidate expressions for substitutions.

Using the resulting DFG, FormOpt replaces three types of string expressions: string literals and constructors, conversion/coercion of any value to a string (for export), and conversion/coercion of any string to a primitive (for import). To illustrate this, Figure 8 shows (a) two potential implementations of an export function, (b) the code after string replacement, and (c) the accumulated values in the internal state of the `AStrings` after one iteration of the loop.

Algorithm 1 summarizes the replacement process. On lines 1-2, the algorithm executes each test and identifies the relevant file IO call sites. On line 4, it uses these sites to construct a DFG. For each expression e that converts to or from a string format, it replaces e with a corresponding `AString` operation (lines 5-14). Lines 6-9 target expressions relevant for data export; for example, a string literal v is replaced with an augmented instance `AString(v)`. To support efficient imports, the algorithm performs a similar replacement for strings converted to primitive values (lines 10-14).

As in before, PipeGen verifies the correctness of the modifications described above by executing the specified unit tests in conjunction with the verification proxy. PipeGen turns off this optimization if one or more unit tests fail following the modifications made by the FormOpt component in string decoration mode (we have not encountered such cases in our experiments).

5.2 Using External Libraries

As mentioned, many DBMSs use external libraries to serialize data. Handling such engines require a custom subtype to be implemented for each external library (of which there are a few that are commonly used). Under this mode, FormOpt replaces instantiations of a given formatting library with a PipeGen-aware subtype that tries to avoid the overhead associated with strings and delimiters. For example, whenever the DBMS invokes a method that builds or parses the text format, PipeGen instead internally constructs or produces a binary representation. When the resulting text fragment is converted to string form, the PipeGen-aware subtype generates an `AString` that contains the binary representation as its internal state. During import, the library subtype recognizes that it is interacting with the data pipe type (q.v. Section 4.1) and directly consumes the intermediate binary representation. This allows the library subtype to construct an efficient internal representation of the input.

As before, FormOpt must identify only those locations where a library is used for import and export. Our approach for doing so – using unit tests and DFGs – is similar to that of string decoration. For example, if a user specifies JSON when invoking PipeGen, FormOpt will examine the source code of the DBMS for instantiations of supported JSON libraries. Using the resulting DFG, FormOpt

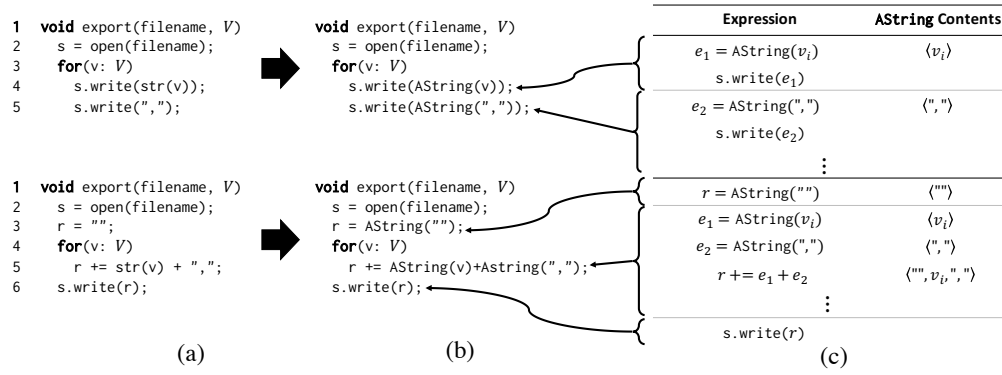


Figure 8: (a) Two ways to implement CSV data export (inserting newlines instead of comma on the last value is omitted for clarity); (b) after string replacement of literals and conversions; (c) accumulated values in the AString instances after one loop iteration.

replaces string literals, constructors, and conversion/coercion expressions in a manner identical to that discussed above. Additionally, FormOpt replaces the instantiation of the library itself with an augmented variant along with any writer or stream interfaces that the library exposes. For example, consider the following simplified version of the Spark JSON export function:

```

String toJSON(RDD<String> rdd) {
  Writer w = new CharArrayWriter();
  JsonGenerator g = new JsonGenerator(w);
  foreach(Object e: rdd) { generateJSON(g, e); }
  return w.toString(); }

```

Our transformations produce the following modified variant:

```

String toJSON(RDD<String> rdd) {
  Writer w = new AWriter(new CharArrayWriter());
  JsonGenerator g =
    new AJsonGenerator(new JsonGenerator(w));
  foreach(Object e: rdd) { generateJSON(g, e); }
  return w.toString(); } // an AString!

```

As in string decoration, FormOpt disables library call replacement if the generated code does not pass all unit test cases. If string decoration also fails to pass the tests, then PipeGen only generates the basic data pipe as discussed in Section 4.

5.3 Intermediate Format

Using AStrings resolves the first two sources of overhead introduced above (encoding of numeric types and conversion/parsing overhead). In this section, we introduce optimizations designed to eliminate delimiters and avoid redundant metadata. These optimizations are implemented inside PipeGen's data pipe type.

5.3.1 Delimiter Inference and Removal

Text-oriented formats such as CSV and JSON include delimiters that separate attributes and denote the start and end of composite types. In some cases these delimiters are fixed in advance; for example, square brackets are used in JSON to indicate an array. However, default delimiters often vary on a per-system basis. This is common under CSV, where some systems default to a non-comma delimiter (e.g., Hadoop uses tab separation by default) or allow the delimiter to be specified by the user (e.g., Derby). To eliminate delimiters, FormOpt needs to first infer them. FormOpt does so by first running the provided unit tests. During the execution of each test, FormOpt counts the length-one strings within the array and identifies which character is most likely to be the delimiter. For example, the array `[1, "l", "a,b", "\n"]` contains exactly one length-one string ("l"), and FormOpt concludes that this is most likely to be the delimiter. The input `[1, "l", "a", "\n"]` is ambiguous, since both

"l" and "a" appear with equal frequency. In this case, FormOpt applies, in order, the following tie-breaking heuristics: (i) prefer non-alphanumeric delimiters, and (ii) prefer earlier delimiters. Under both heuristics, FormOpt would select "l" as the final delimiter.

Note that should FormOpt infer an incorrect delimiter, invalid data will be transmitted to the remote DBMS. In the previous example, if FormOpt's selection of "l" was invalid and the character "a" was actually the correct delimiter, it would incorrectly transmit the tuple (1, "a") instead of the correct value ("1l", ""). More importantly, this is likely to cause the unit tests to fail as discussed in Section 4. This results in FormOpt disabling the optimization until the unit tests were extended to fully disambiguate the inference.

5.3.2 Redundant Metadata Removal

More complex text formats such as JSON may not require the delimiter inference described above, but instead serialize complex composite types such as arrays and dictionaries. When producing/consuming JSON or a similar textual format, the composite types produced by a DBMS often contain values that are highly redundant. For example, consider the following document produced by the Spark toJSON method:

```

{"column1": 1, "column2": "value1"}
{"column1": 2, "column2": "value2"}
{"column1": 3, "column2": "value3"}

```

When such JSON documents are moved between systems, the repeated column names greatly increase the size of the intermediate transfer. To avoid this overhead, FormOpt modifies the format of the intermediate data to transmit exactly once the set of keys associated with an array of dictionaries. In the above example, FormOpt would transmit the column names `["column1", "column2"]` as a *key header*, and then the values `[(1, "value1"), ...]` as a sequence of pairs. When importing, FormOpt reverses this process to produce the original JSON document(s).

The logic for this transformation is embedded into the JSON state machine (a subcomponent of the data pipe type) that is used to consume the AString array. When FormOpt transitions into the key state for the first dictionary in an array, it accumulates that key in the key header. Once the dictionary has been fully examined, PipeGen transmits the key header to the remote DBMS. Subsequent dictionaries in that array are transmitted without keys, so long as they are identical to the initial dictionary. While this approach may be extended to nested JSON documents, our prototype currently only optimizes top-level dictionaries.

If a new key is encountered in some subsequent dictionary after the key header has been transmitted, FormOpt adopts one of two

strategies. First, if the keys from the new dictionary are a superset of those found in the key header, FormOpt appends the new key to the existing key header. This addresses the common case where the set of exported keys was not complete due to, for example, a missing value in the initial exported dictionary.

A second case occurs when the keys associated with a new dictionary are disjoint from those in the key header. This might occur during export from a schema-free DBMS, where exported elements have widely varying formats. In this case, FormOpt disables the optimization for the current dictionary and does not remove keys during its transmission.

5.4 Column Orientation & Compression

DBMSs that output text-oriented formats generally do so in a row-oriented manner. For example, a Spark RDD containing n elements that is exported to CSV or JSON will generate n lines, each containing one element in the RDD. This is also true in the other systems we evaluate, for both JSON and CSV formats. However, once FormOpt produces an efficient data representation, we no longer need to transmit data in row-major form. For example, the data pipe type can accumulate blocks of exported data and transform it to column-major form to improve transfer performance. Indeed, recent work on column-oriented DBMSs suggests that some of the benefits (e.g., compacting/compression, improved IO efficiency) [38] may also improve performance for data transfer between DBMSs.

After examining several formats for the wire representation of our data (see Section 7.3) we settled on Apache Arrow as the data structure we transmit, since it performs the best. To maximize performance, our prototype accumulates blocks of rows in memory, pivots them into column-major form by embedding them into a set of Arrow buffers, and transmits these buffers to the destination DBMS. The receiving DBMS reverses this process.

6. IMPLEMENTATION

We have implemented a prototype of PipeGen in Java. The generated data pipes currently support DBMSs that make use of the local file system or HDFS for data import and export.

6.1 File IO Redirection

The file IO redirector in the current prototype targets `FileInputStream` and `FileOutputStream` as the relevant file system calls to be modified. In addition to our concrete implementation of the data pipe class (`DataPipeInput/OutputStream`), we created augmented versions of the following Java classes:

- `StringBuilder/Buffer`. We introduced an array analogous to the one found in `AString`.
- `Output/InputStreamWriter`. These classes were modified to interact with the `DataInput/OutputStream` classes and contained overloads for string IO.
- `BufferedOutput/InputStream`. These classes were augmented to detect when the underlying stream was a data pipe class, and if so buffering was omitted.
- `org.apache.hadoop.io.Text`. We augmented this class with an object array in the same way as the `AString` class. We needed to do so because the Hadoop-based systems uses this class similar to the ordinary Java strings.
- `org.apache.hadoop.hdfs.DFSInput/OutputStream`.

We produced HDFS-specific data pipe classes for these classes. Implementation was analogous to that of `DataPipeInput/OutputStream` classes.

- `java.sql.ResultSet`. We replaced the `getString` methods with a version aware of our `AString` class.

Our implementation of `DataPipeInputStream` supports seeks within one data row. This is needed to support HDFS; when the HDFS client opens a file, it performs a small read/rewind in order to determine whether the file being read is a Hadoop sequence file.

Two systems that we evaluated disallowed URIs with a custom scheme (e.g., `dbms://A`) as an import/export target (Derby and Myria). For these systems, our prototype supports the ability to specify the reserved filename template in a configuration file. In each case we substituted an alternative of the form `"\tmp__reserved__[Name]"`. These same systems perform an explicit check for the existence of the reserved filename prior to proceeding with an import. Similarly, DBMSs that import from HDFS calculate split points by examining files before beginning the import process. To protect against failures in each of these cases, PipeGen automatically creates stub instances of the reserved filename on the file system during startup.

6.2 Augmented String Implementation

We previously defined `AString` to be a subclass of `java.lang.String`, which is declared to be `final`. We work around this issue by replacing the standard `String` class with a non-final version via dynamic code loading.

For performance reasons the implementation of `AString` in Java maintains its array of values as a flat byte array. These byte arrays are preallocated on startup and managed internally by `AString`. For operations that may not be performed on the raw array (e.g., substring), `AString` falls back to a materialized string representation.

Finally, since Java does not support operator overloading, PipeGen rewrites all string concatenation in the source code using functional forms. For instance, for `a+b` where `a` and `b` are both `java.lang.String`, PipeGen rewrites the expression into `new AString(a).concat(new AString(b))` instead.

7. EVALUATION

We evaluate PipeGen using benchmarks that show data transfer times between five Java DBMSs: Myria [20], Spark 1.5.2 [44], Giraph 1.0.0 [7], Hadoop 2.7.1 [5], and Derby 10.12.1.1 [4]. The first set of experiments examines the overall performance differences between the data pipes generated by PipeGen and importing/exporting data through the file system (Section 7.1). Next, we analyze the performance gains from each of our optimizations (Section 7.2). We then evaluate the impact of different data formats transferred between DBMSs (Section 7.3), along with the compression method used during transport (Section 7.4). Finally, we examine the number of modifications done by each step during data pipe compilation (Section 7.5).

Unless otherwise specified, experiments utilize a 16-node cluster of `m4.2xlarge` instances in the Amazon Elastic Compute Cloud. Each node has 4 virtual CPUs, 16 GiB of RAM, and an attached 1TB standard elastic block storage device. We deploy the most recent stable release of each DBMS running under OpenJDK 1.8. Except for Derby, which is a single-node DBMS, we deploy each system across all 16 instances using YARN [42]. For each pair of DBMSs, we colocate workers on each node and assign each YARN container two cores and 8 GiB of RAM.

With the exception of Figure 10, the experiments in this section all involve the transfer of n elements with a schema having a unique integer key in the range $[0, n]$ followed by three (integer $\in [0, n]$, double) pairs. Each 8-byte double was sampled from a standard normal distribution. For Giraph, we interpreted the values as a graph having n weighted vertices each with three random directed edges weighted by the following double.

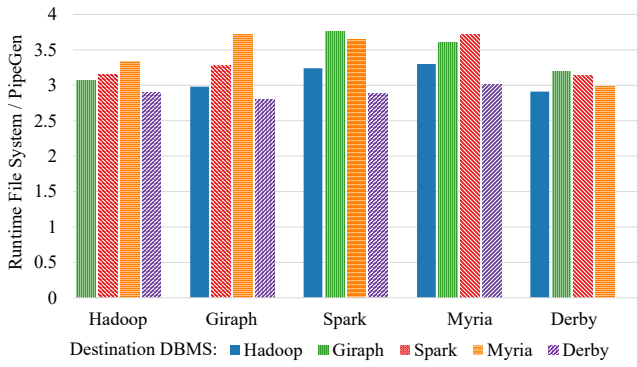


Figure 9: Total speedup between file system and PipeGen for 10^9 elements. Transfer occurred from a source DBMS (x -axis) to a destination DBMS (bar color/pattern) using the CSV format. The number of workers/tasks was fixed at 16.

# Workers	1	4	8	16
Speedup	3.1	3.7	3.5	3.7

Table 1: Overall speedup (file system / PipeGen runtime) from Myria to Spark for $4 \cdot 10^8$ elements when varying the number of workers and tasks involved in the transfer.

7.1 Paired Transfer

We first evaluate the overall benefit of our approach for different combinations of systems. Figure 9 shows the total transfer time between a source and destination DBMS using (i) an export/import through the file system based on functionality provided by the original DBMS, and (ii) an export/import using the PipeGen-generated data pipes. For this experiment, we transfer 10^9 elements, fix the number of workers/tasks at 16 for each DBMS, and enable all optimizations. Since CSV is the only common format supported by all DBMSs, file system transfers use this format and the PipeGen data pipes are generated from CSV export and import code.

As the results show, data pipes significantly outperform their file system-oriented counterparts. For this transfer size, the average speedup over all DBMSs is $3.2\times$, with maximum speedup up to $3.8\times$. This speedup is approximately the same across all transfer sizes and pairs of DBMSs. As shown in Table 1, this speedup is also similar across various cluster sizes.

This result emphasizes the impact that PipeGen can have on hybrid data analytics: without writing a single line of code, a user can get access to 20 optimized data pipes and speed up data transfers between any combination of the five systems tested by $3.2\times$ on average. PipeGen produces this benefit automatically without requiring that individual system developers agree on an efficient common data format. Each system developer only needs to implement CSV export and import methods.

The magnitude of the benefit does depend on the types of data transferred. Figure 10 shows the speedup between the file system and PipeGen for $4 \cdot 10^8$ elements of various data types. As the figure shows, the transfer performance for fixed-width primitives perform significantly better than string transfers, due to the smaller amount of data transferred when using AStrings. While strings do not benefit from our optimizations, they still benefit from avoiding serializing to the file system during data transfer.

7.2 Optimizations

We drill down into the different components of the speedup afforded by PipeGen’s data pipes. In this section, we evaluate the

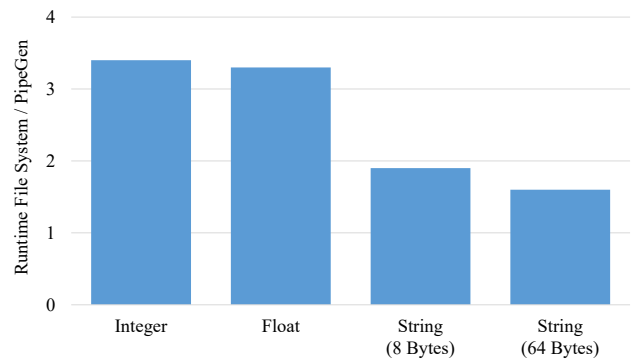


Figure 10: Overall speedup between file system and PipeGen transfer for different data types and sizes. Each transfer moves $4 \cdot 10^8$ elements of the given data type.

performance benefits of FormOpt’s optimizations, which convert text-formatted data to binary after removing delimiters and meta-data.

7.2.1 String Decorations

We first evaluate the optimizations due to the use of augmented strings as described in Section 5.1. Figure 11 shows the performance of an export between Myria and Giraph using this mode and individual optimizations. For this pair, FormOpt’s optimizations are responsible for approximately one third of the total runtime benefit beyond what IORedirect already provides.

To assess the performance benefits of avoiding both text-encoding and delimiters, we compare the performance against a manually-constructed data pipe that implements only these two optimizations. To produce the manually constructed pipes, we modify each DBMS to directly transmit/receive results to/from a network socket and eliminate all intermediate logic related to text-encoding that might degrade performance.

In our experiment, we transfer data in both directions and measure the total runtime. Overall, the PipeGen-generated data pipes perform closely to their manually-optimized counterparts. Transferring from Myria to Giraph is slower due to the implementation of the Giraph import, where Giraph materializes AString instances into character strings, and processes characters from the materialized string to escape them if needed.

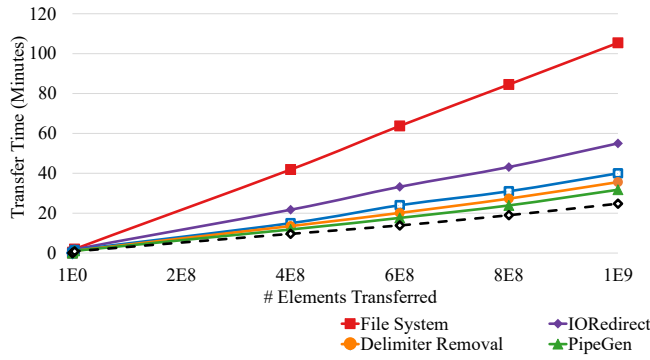
7.2.2 Library Extensions

We generate a library extension implementation for the Jackson JSON library and evaluate its performance under the library extension mode of FormOpt. Interestingly, for the pairs of DBMSs that we examine, most do not support mutually-compatible exchange using JSON as an intermediate format. For example, Myria produces a single JSON document, Spark and Giraph both expect a document-per-line, and Derby does not natively support bulk import and export of JSON at all.

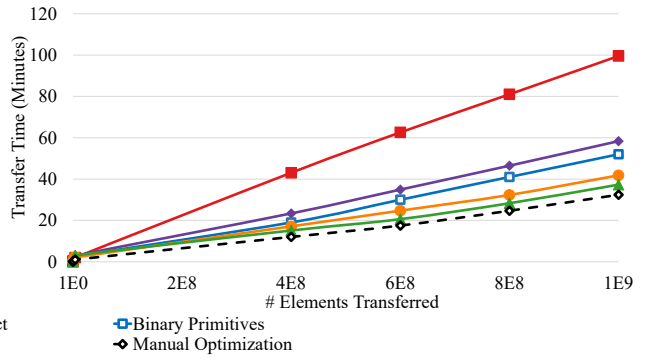
Figure 12 shows the performance of using library extensions with Jackson between Spark and Giraph. As before, we use a mutually-compatible JSON adjacency-list format for the schema of transmitted data. We find that the relative performance benefit closely matches that of the string decorations.

7.3 Intermediate Format

Once FormOpt captures the transferred data in AString we can use any intermediate format to transfer the data between DBMSs. In this section, we show that, as expected, the choice of that intermediate format significantly impacts performance, with recent



(a) Giraph→Myria DBMS



(b) Myria DBMS→Giraph

Figure 11: Transfer time between the Myria and Giraph using PipeGen and a manually-optimized variant. In (a) we export tuples from Myria and import them as vertices in Giraph. In (b) we reverse the direction of transfer. We show as baseline transfer through the file system. We then activate PipeGen optimizations as follows. First, we apply the IORedirect component. Next, we transmit fixed-width values in binary form. We then activate delimiter removal. The PipeGen series shows all optimizations, which additionally include column pivoting.

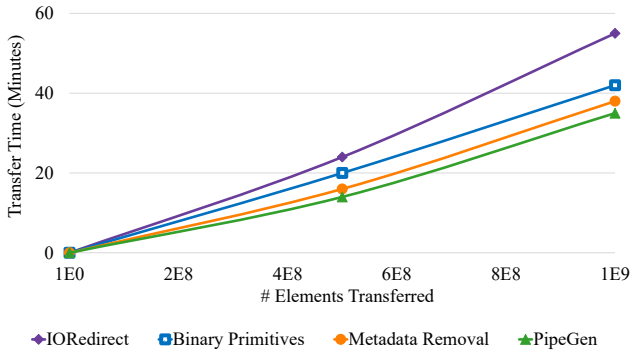


Figure 12: Total runtime of a transfer from Spark to Giraph using the library extension mode for the Jackson JSON library. We show a baseline application of only the IORedirect component. We then transmit fixed-width values in binary form. Next, we activate metadata removal (i.e., repeated column names and delimiters). The PipeGen series shows all optimizations (i.e., column pivoting) activated.

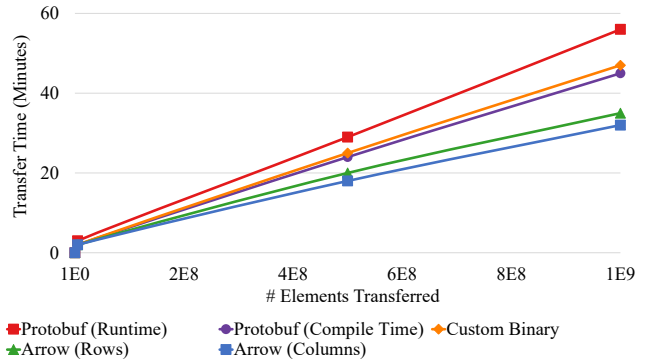


Figure 13: Transfer performance by intermediate format between Hadoop and Spark. Message templates for protocol buffers were generated both at compile time and dynamically at runtime.

formats outperforming older ones. This observation is important as a key contribution of PipeGen is to free developers from the need to add new data import and export code to the DBMS every time a new data format becomes available.

Our experiments include two third-party formats: protocol buffers [32] and Arrow [6]. We also evaluate the basic custom format from the previous section, which transmits schema information as a header, fixed width values in binary form, and uses length-prefixing for strings. We examine protocol buffers using a version where message templates are fixed at compile time and another where they are dynamically constructed at runtime. Figure 13 shows the results. Protocol buffers, depending on whether message formats are statically or dynamically generated, perform approximately as well as our custom binary format. The recent data format, Arrow, offers a substantial boost in performance, due primarily to its optimized layout, efficient allocation process, and optimized iteration [6]. Additionally, the column-oriented format offers a further modest advantage over its row-oriented counterpart.

It is clear that Apache Arrow yields the highest performance as an intermediate format. Since we preallocate a buffer for use dur-

ing block-oriented transfer necessary for pivoting the data into a columnar format, it is necessary to reserve an appropriate size for this intermediate buffer. In Figure 14, we illustrate transfer performance between Myria and Giraph for various ArrowBuf sizes. Note that since Arrow is column-oriented, we allocate one buffer for each attribute. As with our previous experiments, this includes 4 four-byte integers and 4 eight-byte doubles. This implies that for a n -element buffer we preallocate eight ArrowBufs for a combined total of $48n$ bytes (plus any bookkeeping overhead required by Arrow’s implementation). As long as the buffer is not too small, the buffer size has only a negligible impact on performance.

7.4 Compression

Orthogonal to the choice of data format and application of optimizations is PipeGen’s ability to compress the data transferred between pairs of DBMSs. Utility of this approach depends on the distance between DBMS workers, with nearby DBMSs being less likely to benefit than distant ones.

Figure 15 shows the performance of transfer using compression (or lack thereof) from Myria to Giraph. We use three compression techniques: run-length encoding, dictionary-based compression (zip), and uncompressed transfer. We separately show transfer performance between colocated workers (Figure 15(a)) and

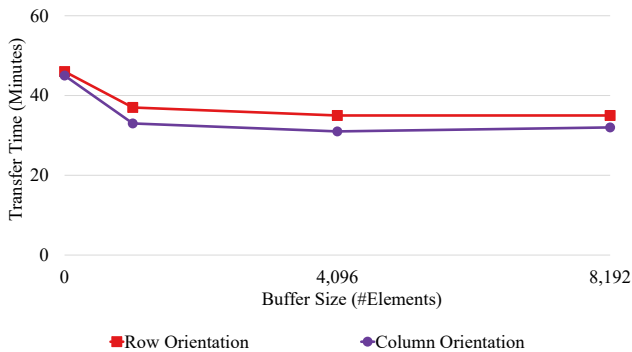


Figure 14: Transfer time from Myria to Giraph for 10^9 elements with varying sizes of Arrow buffers. Each of the column buffers used for the transfer was sized to hold the number of values listed on the x -axis.

workers with a 40ms artificial latency introduced into the network adapter (Figure 15(b)). For colocated workers, we also evaluate transfer performance using shared memory; all other transfers take place over sockets.

For colocated nodes, both compression techniques add modest overhead to the transfer process that yields a net loss in performance. For nodes with higher latency, we show a modest benefit for dictionary-oriented compression. This suggests that this strategy may be beneficial for physically-separated DBMSs.

7.5 Code Modifications

To get a sense of the amount of changes needed to generate the data pipes for each DBMSs, Table 2 lists the number of modifications made by the IORedirect and FormOpt components, in terms of the number of classes and lines of code affected. We also measure the amount of time PipeGen spent to perform the compile-time modifications.

As the results show, the amount of changes is modest across all of the DBMSs we evaluate. In addition, the total number of modifications is small for the IORedirect component. This shows that DBMS implementations rarely open an import or export file at more than one call site. The FormOpt component performs more modifications, with library extension requiring fewer changes than string decoration. Even with string decoration, for the DBMSs that we evaluate, primitive values are converted to/from a string in close proximity in the code where they are written/read. This reduced the number of modifications required for this optimization.

8. RELATED WORK

There have been a number of systems that implement hybrid data analytics using an orchestrator-based architecture over various constituent DBMSs. In such systems, each constituent must be manually integrated into the hybrid engine. This architecture was present in the earliest heterogeneous systems such as COSYS [1], HD-DBMS [10], MDAS [12], InterViso [40], and Tukwila [21], and is common in more recent systems such as the Cyclops [26], Polystore [13], Estocada [9], and Musketeer [17]. Our data pipes complement these systems by providing an efficient mechanism to transfer large amounts of data between them.

Data exchange is another related topic that also involves transferring of data across heterogeneous systems. However, while there has been much work on generating robust mappings between schemata across systems [15], optimization has focused primarily on inference performance [35, 19], or does not address data shipping performance [33]. In contrast, PipeGen focuses on optimizing

data transfer performance, and assumes that the user or a query optimizer can generate query plans to reconcile the different schemata across DBMSs.

PipeGen automatically embeds data pipes into a selected DBMS. Some previous work in the area of data integration has explored the manual embedding of data pipes for transferring data across heterogeneous systems. For example, Rusinkiewicz et al. proposed a common inter-DBMS format mediated via STUB operators [34]. These operators are similar to data pipes, but are manually generated and do not address direct transfer between DBMSs. Additionally, prior work has explored the automatic generation of data transformation operators that are similar to data pipes. For example, Mendell et al. used code generation under a general-purpose streaming engine to support streaming XML processing [28]. Similar tools target automatic parsing of semi-structured data for ad hoc processing [16]. While these efforts shares some commonality with PipeGen, they target a single data model and do not address data transfer performance between DBMSs.

There are many potential formats (e.g., protocol buffers [32], Parquet [31], Thrift [41], Avro [8], Arrow [6]) that can be used to move data between two DBMSs. Since each of these requires manual effort in order to be supported by a DBMS, PipeGen’s ability to switch between formats as they evolve allows users to capture performance benefits with reduced effort.

Finally, PipeGen’s IORedirect component redirects specific file system calls to network socket operations instead. This redirection is similar to previous work in the systems research community. For example, the CDE software distribution tool intercepts file system calls using ptrace and redirects IO to an alternate location [18]; similar work by Spillane et al. used system call redirection for rapid prototyping [36]. This approach is also common in the security community for applications ranging from sandboxing [43] to information flow analysis [45].

9. CONCLUSION

In this paper we described PipeGen, a tool that automatically generates data pipes between DBMSs for efficient data transfer. Observing that most DBMSs support importing from and exporting to a common data format, PipeGen makes use of existing test cases and program analysis techniques to create data pipes from existing source code, and furthermore optimizes the generated implementation by removing several inefficiencies in data encoding and unnecessary data serialization.

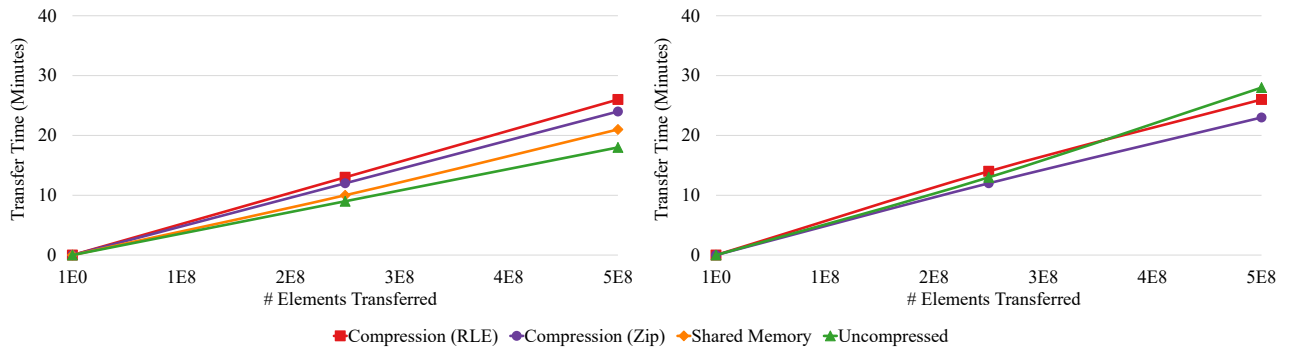
We have implemented a prototype of PipeGen and evaluated it by automatically generating data pipes across a variety of DBMSs, including relational and graph-based engines. Our experiments show that the PipeGen-generated data pipes enables efficient hybrid data analytics by outperforming the traditional way of transferring data via the file system by up to $3.8\times$.

Acknowledgements

This work is supported in part by the National Science Foundation through NSF grants IIS-1247469 and IIS-1110370, and gifts from the Intel Science and Technology Center for Big Data, Amazon, and Facebook.

10. REFERENCES

- [1] M. Adiba and D. Portal. A cooperation system for heterogeneous data base management systems. *Information Systems*, 3(3):209–215, 1978.
- [2] L. Andersen. Jdbc 4.2. Technical Report JSR 221, Oracle, March 2014.
- [3] Apache Commons CSV. <https://commons.apache.org/proper/commons-csv/>.
- [4] Apache Software Foundation. Derby. <https://db.apache.org/derby>.
- [5] Apache Software Foundation. Hadoop. <https://hadoop.apache.org>.
- [6] Apache arrow. <https://arrow.apache.org/>.
- [7] C. Avery. Giraph: Large-scale graph processing infrastructure on hadoop. *Proceedings of the Hadoop Summit, Santa Clara*, 2011.



(a) Myria→ Giraph, colocated workers

(b) Myria→ Giraph, 40ms latency

Figure 15: Transfer time between Myria and Giraph using compression. In (a) we transfer between workers colocated in the same physical node, while in (b) we transfer between workers with 40ms latency artificially introduced. For RLE, zip, and uncompressed formats we transfer data using a socket. For colocated nodes, we also transfer data uncompressed using shared memory.

Mode	DBMS	Modification Time (seconds)	IORedirect		FormOpt	
			#Classes	LOC	#Classes	LOC
String Decoration	Hadoop	245	3	6	6	36
	Myria	160	2	8	5	54
	Giraph	223	2	9	4	47
	Spark	187	5	18	8	38
	Derby	130	2	5	2	67
Library Extension	Spark	178	5	18	2	6

Table 2: For each of the DBMSs we evaluated, we show the time required for PipeGen to modify each system. In addition, for each of the two phases, we also show the number of classes and lines of code modified by each component.

- [8] Apache Avro. <https://avro.apache.org/>.
- [9] F. Bugiotti, D. Bursztyn, A. Deutsch, I. Ileana, and I. Manolescu. Toward scalable hybrid stores. In *SEBD*, 2015.
- [10] A. F. Cardenas. Heterogeneous distributed database management: The hd-dbms. *Proceedings of the IEEE*, 75(5):588–600, 1987.
- [11] Dato, Inc. Graphlab integration with spark open source release. <http://blog.dato.com/graphlab-integration-with-spark>, 2015.
- [12] B. C. Desai and R. Pollock. Mdas: heterogeneous distributed database management system. *Information and Software Technology*, 34(1):28–42, 1992.
- [13] D. J. DeWitt, A. Halverson, R. Nehme, S. Shankar, J. Aguilar-Saborit, A. Avanes, M. Flaszka, and J. Gramling. Split query processing in polybase. In *SIGMOD*, pages 1255–1266. ACM, 2013.
- [14] A. Elmore, J. Duggan, M. Stonebraker, M. Balazinska, U. Cetintemel, V. Gadepally, J. Heer, B. Howe, J. Kepner, T. Kraska, et al. A demonstration of the bigdaw polystore system. *VLDB*, 8(12):1908–1911, 2015.
- [15] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: semantics and query answering. *Theoretical Computer Science*, 336(1):89–124, 2005.
- [16] K. Fisher, D. Walker, K. Q. Zhu, and P. White. From dirt to shovels: Fully automatic tool generation from ad hoc data. In *POPL*, page 421. ACM, 2008.
- [17] I. Gog, M. Schwarzkopf, N. Crooks, M. P. Grosvenor, A. Clement, and S. Hand. Musketeer: all for one, one for all in data processing systems. In *Proceedings of the Tenth European Conference on Computer Systems*, page 2. ACM, 2015.
- [18] P. J. Guo and D. R. Engler. Cde: Using system call interposition to automatically create portable software packages. In *USENIX ATC*, 2011.
- [19] L. M. Haas, M. A. Hernández, H. Ho, L. Popa, and M. Roth. Clio grows up: from research prototype to industrial tool. In *SIGMOD*, page 805. ACM, 2005.
- [20] D. Halperin, V. T. de Almeida, L. L. Choo, S. Chu, P. Koutris, D. Moritz, J. Ortiz, V. Ruamviboonsuk, J. Wang, A. Whitaker, et al. Demonstration of the myria big data management service. In *SIGMOD*, pages 881–884. ACM, 2014.
- [21] Z. G. Ives, D. Florescu, M. Friedman, A. Levy, and D. S. Weld. An adaptive query execution system for data integration. In *ACM SIGMOD Record*, volume 28, pages 299–310. ACM, 1999.
- [22] Jackson JSON Processor. <http://wiki.fasterxml.com/JacksonHome/>.
- [23] P. Jetley, F. Gioachin, C. Mendes, L. V. Kale, and T. Quinn. Massively parallel cosmological simulations with changa. In *IPDPS*, pages 1–12. IEEE, 2008.
- [24] V. Josifovski, P. Schwarz, L. Haas, and E. Lin. Garlic: a new flavor of federated query processing for db2. In *SIGMOD*, pages 524–532. ACM, 2002.
- [25] A. Knebe, F. R. Pearce, H. Lux, Y. Ascasisar, P. Behrooz, J. Casado, C. C. Moran, J. Diemand, K. Dolag, R. Dominguez-Tenreiro, et al. Structure finding in cosmological simulations: the state of affairs. *MNRAS*, 435(2):1618, 2013.
- [26] H. Lim, Y. Han, and S. Babu. How to fit when no one size fits. In *CIDR*, volume 4, page 35, 2013.
- [27] F. Lin and W. W. Cohen. Power iteration clustering. In *ICML*, page 655, 2010.
- [28] M. Mendell, H. Nasgaard, E. Bouillet, M. Hirzel, and B. Gedik. Extending a general-purpose streaming system for xml. In *EDBT*, page 534. ACM, 2012.
- [29] Myria: Big Data management as a Cloud service. <http://myria.cs.washington.edu/>.
- [30] Microsoft Open Database Connectivity (ODBC). <https://msdn.microsoft.com/en-us/library/ms710252>.
- [31] Apache Parquet. <https://parquet.apache.org/>.
- [32] Protocol Buffers. <https://developers.google.com/protocol-buffers/>.
- [33] T. Risch, V. Josifovski, and T. Katchaounov. Functional data integration in a distributed mediator system. In *The Functional Approach to Data Management*, pages 211–238. Springer, 2004.
- [34] M. Rusinkiewicz, K. Loa, and A. K. Elmagarmid. Distributed operation language for specification and processing of multidatabase applications. 1988.
- [35] K. Saleem, Z. Bellahsene, and E. Hunt. Porsche: Performance oriented schema mediation. *Information Systems*, 33(7):637–657, 2008.
- [36] R. P. Spillane, C. P. Wright, G. Sivathanu, and E. Zadok. Rapid file system development using ptrace. In *ExpCS*, page 22. ACM, 2007.
- [37] M. Stonebraker. ACM SIGMOD blog: The case for polystores. <http://wp.sigmod.org/?p=1629>.
- [38] M. Stonebraker, D. J. Abadi, A. Batkin, X. Chen, M. Cherniack, M. Ferreira, E. Lau, A. Lin, S. Madden, E. O’Neil, et al. C-store: a column-oriented dbms. In *VLDB*, pages 553–564. VLDB Endowment, 2005.
- [39] M. Stonebraker, P. Brown, A. Poliakov, and S. Raman. The architecture of scidb. In *SSDBM*, pages 1–16. Springer, 2011.
- [40] M. Templeton, H. Henley, E. Maros, and D. J. Van Buer. Interviso: Dealing with the complexity of federated database access. *VLDB*, 4(2):287–318, 1995.
- [41] Apache Thrift. <https://thrift.apache.org/>.
- [42] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, et al. Apache hadoop yarn: Yet another resource negotiator. In *SOC*, page 5. ACM, 2013.
- [43] D. Wagner, I. Goldberg, and R. Thomas. A secure environment for untrusted helper applications. In *Proc. of the 6th USENIX Unix Security Symp.*, 1996.
- [44] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: cluster computing with working sets. In *HotCloud*, pages 10–10, 2010.
- [45] N. Zeldovich, S. Boyd-Wickizer, E. Kohler, and D. Mazières. Making information flow explicit in histar. In *OSDI*, pages 263–278. USENIX Association, 2006.